基于矩阵加权关联模式的印尼中跨语言信息检索模型*

黄名选
(广西跨境电商智能信息处理重点实验室培育基地(广西财经学院) 南宁 530003)
(广西财经学院计算机系 南宁 530003)

摘要:【目的】针对跨语言信息检索存在的查询漂移问题, 提出一种融合用户点击下载行为与矩阵加权关联模式挖掘的印尼中跨语言信息检索模型。【方法】将矩阵加权关联模式挖掘、查询扩展以及用户点击下载行为集成应用到印尼中跨语言信息检索模型, 给出模型实现的关键技术, 即面向跨语言信息检索的矩阵加权关联模式挖掘算法、跨语言查询扩展模型以及印尼中跨语言信息检索算法。【结果】在 NTCIR-5 CLIR 数据集上的实验结果表明, 该检索模型的 R_prec、p@10 和 p@20 值均达到单语言检索基准的 60%以上, 比跨语言检索基准提高 37%以上, 比现有基于伪相关反馈的跨语言检索算法提高 28%以上。【局限】该模型实验在基于向量空间模型的跨语言检索系统中进行, 需要探讨和研究在实际搜索引擎中的具体应用。【结论】该模型能有效地减少跨语言检索中的查询漂移问题, 提高和改善印尼中跨语言检索性能, 对长查询的检索效果更好, 有较好的实际应用价值。

关键词: 点击行为 关联模式挖掘 印尼中跨语言检索模型 跨语言信息检索 矩阵加权关联规则
分类号: TP311

1 引 言

跨语言信息检索指的是以一种语言检索出其他语言的信息资源的技术。印尼中跨语言信息检索指的是使用印尼语检索中文文档, 其中, 用于查询的印尼语言称为源语言(Source Language, SL), 中文称为目标语言(Target Language, TL)。世界各地学者从不同的角度和方向对跨语言信息检索模型与算法进行了深入探讨和研究, 取得了丰富的理论成果, 然而, 跨语言信息检索研究所存在的问题还没有完全解决, 该领域亟待解决和关注度比较高的问题之一是跨语言信息检索比单语言检索面临更为严重的词不匹配和主题漂移问题, 这些问题常常导致跨语言检索性能低下。针对这些问题, 近年来, 基于查询扩展的跨语言信息检索研究得到了更多的关注和讨论, 其研究主要集中在基于相关反馈[1-6]、潜在语义[7-10]、语言模型[11]和主题模型[12-16]等跨语言信息检索研究, 其语言对象以英语为主, 大多都是研究英语和其他语言的跨语言检索问题。

基于相关反馈的跨语言信息检索即利用跨语言初检结果的前列文档作为跨语言查询扩展词项的来源实现查询扩展, 然后再次检索文档。其典型算法是 Gao 等[1]提出的两步伪相关反馈法。吴丹等[2]在此基础上对基于伪相关反馈的跨语言查询扩展进行深入研究, 通过伪相关反馈实验比较 4 种跨语言信息检索查询翻译优化技术[3], 取得较好的研究成果。近年来, Chinnakotla 等[4]提出使用与查询不同的辅助语言材料改善跨语言

通讯作者: 黄名选, ORCID: 0000-0002-5942-5295, E-mail: mingxh05@163.com。
*本文系国家自然科学基金项目"面向东盟国家语言的基于完全加权正负模式挖掘的跨语言查询扩展研究"(项目编号: 61262028)、广西财经学院信息与统计学院开放性课题"基于矩阵加权关联模式挖掘的越汉英跨语言信息检索研究"(项目编号: 2015XK01)和广西财经学院2016年度应用统计硕士专业学位点学术研究项目"基于完全加权关联模式挖掘的中英跨语言伪相关反馈扩展研究"(项目编号: 2016TJYB05)的研究成果之一。

伪相关反馈扩展性能，以提高跨语言检索效率。Parton 等[5]将机器学习引入跨语言相关反馈扩展领域，Lee 等[6]针对博客或论坛等非正式文本，提出一种新的伪相关反馈扩展技术改善跨语言检索性能，都取得了良好的实验结果。

基于潜在语义的跨语言信息检索即利用潜在语义分析技术建立不同语言之间的对应关系，从中发现与原查询相关的目标语言特征词，实现跨语言查询扩展，改善跨语言信息检索性能。其典型算法是闭剑婷等[7]提出的通过潜在语义分析的跨语言查询扩展改善跨语言检索性能。魏露等[8]对文献[7]进行改进，通过结合奇异值分解和非负矩阵分解法建立双语空间，改善跨语言检索性能。此后，宁健等[9]通过改进的潜在语义分析实现双语摘要跨语言检索，取得了较好的实验结果。罗远胜等[10]通过双语平行语料库构造每种语言的潜在语义空间提高和改善跨语言检索性能，实验结果表明上述方法是有效的。

基于语言模型、主题模型的跨语言信息检索研究也开始活跃起来。Rahimi 等[11]利用语言模型框架实现跨语言查询扩展，提高了跨语言检索性能。Ganguly 等[12]利用潜在主题对跨语言相关性模型进行改进，以帮助改善目标语言检索效果。此后，Wang 等[13-16]对基于主题模型的跨语言信息检索进行了深入研究，先后提出基于潜在狄利克雷分配(Latent Dirichlet Allocation, LDA)主题模型的跨语言伪相关反馈扩展[13-14]、基于双语主题的跨语言伪相关反馈[15]，以及基于弱相关主题对齐的跨语言伪相关反馈扩展[16]，理论分析与实验结果均表明上述方法是有效性的。

从相关文献报道可以看出，面向东盟国家语言的跨语言信息检索研究还鲜有报道。自中国南宁市作为中国–东盟博览会永久举办地以来，中国与东盟国家的政治、经济、文化等往来更加频繁和密切，面向东盟国家语言的跨语言信息检索和跨语言信息服务研究显得更加迫切，其重要性日益凸显。为此，本文在上述研究成果的基础上，开展面向东盟国家语言的跨语言信息检索研究。以印尼语和汉语为研究对象，将矩阵加权关联规则挖掘技术、用户点击行为与查询扩展等技术集成应用于印尼中跨语言信息检索，提出基于矩阵加权关联模式挖掘的印尼中跨语言信息检索模型及实现该模型的关键技术，即面向跨语言信息检索的矩阵加权关联模式挖掘算法、跨语言查询扩展模型以及印尼中跨语言信息检索算法。

2 基于矩阵加权关联模式挖掘的印尼中跨语言信息检索模型

2.1 设计思想

基于矩阵加权关联模式挖掘的印尼中跨语言信息检索模型基本思想是：首先将印尼语查询通过机器翻译系统译为中文查询，提交给搜索引擎实现跨语言检索中文文档，通过用户对初检文档浏览点击下载行为确认该篇文档为用户反馈初检相关文档，然后应用本文提出的面向跨语言信息检索的矩阵加权关联模式挖掘技术从初检相关文档中挖掘与中文查询相关的扩展词实现跨语言译后扩展，扩展词与原查询组合再次提交给搜索引擎检索，将检索结果经机器翻译为印尼语文档返回给用户。

2.2 模型结构图及其模块功能

根据上述设计思想，给出了基于矩阵加权关联模式挖掘的印尼中跨语言信息检索模型结构图，如图 1 所示。该模型由机器翻译模块、搜索引擎模块、用户点击行为相关反馈提取模块、文档预处理模块、面向印尼中跨语言检索的矩阵加权关联规则挖掘模块、跨语言查询扩展词生成模块、跨语言查询扩展实现模块和最终结果显示模块等 8 个模块和 3 个数据库组成，即初检相关文档数据库、矩阵加权关联规则库和扩展词库。

(1) 机器翻译模块：使用必应机器翻译接口，即 Microsoft Translator API①，主要功能是将用户提交的印尼语查询翻译为中文查询，以及将最终检索结果的中文文档翻译为印尼语文档提交给用户。

(2) 搜索引擎模块：可以使用谷歌或百度等搜索引擎，主要功能是对译后的中文查询在互联网上进行检索，得到跨语言初检结果文档集。

(3) 用户点击行为相关反馈提取模块：捕捉用户浏览初检结果文档集时所产生的文档下载行为，提取用户下载的初检文档构建用户反馈相关文档集。

①https://datamarket.azure.com/dataset/bing/microsofttranslator.

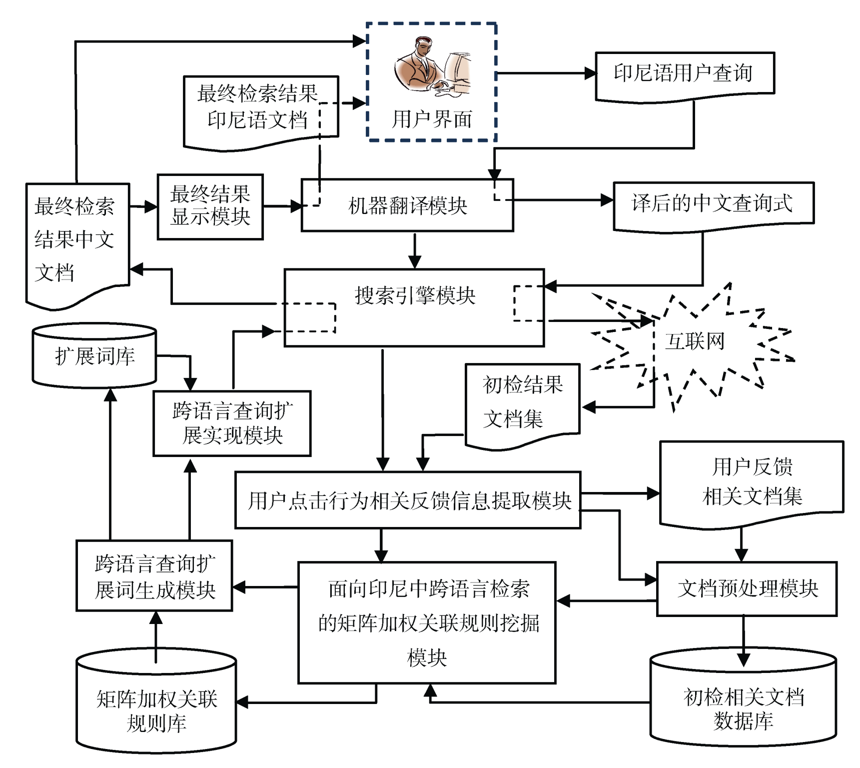


图 1　基于矩阵加权关联模式挖掘的印尼中跨语言信息检索模型

(4) 文档预处理模块：将用户反馈相关文档集进行中文分词、去停用词和提取特征词等预处理，构建用户反馈初检相关文档数据库。

(5) 面向印尼中跨语言检索的矩阵加权关联规则挖掘模块：对上述的用户反馈初检相关文档集进行矩阵加权关联规则挖掘，主要挖掘含有原查询词项的矩阵加权特征词项频繁项集和关联规则模式，构建矩阵加权关联规则库。

(6) 跨语言查询扩展词生成模块：从矩阵加权关联规则库中提取与原查询相关的扩展词，构建扩展词库。

(7) 跨语言查询扩展实现模块：从扩展词库中提取中文扩展词，将扩展词和原查询组合成新查询，再次提交给搜索引擎在互联网中检索，得到最终检索的中文文档。

(8) 最终结果显示模块：将最终检索结果中文文档提交到机器翻译模块翻译为印尼语文档，并将最终检索结果中文文档和印尼语文档返回用户。

2.3　印尼中跨语言信息检索模型关键技术

(1) 面向印尼中跨语言检索的矩阵加权关联规则挖掘

面向印尼中跨语言检索的矩阵加权关联规则挖掘基本思想是：首先通过用户点击行为相关反馈信息提取模块获得印尼中跨语言初检结果，即用户相关反馈目标语言文档集 Doc^{TL}，并由文档预处理模块对 Doc^{TL} 进行预处理，构建用户反馈初检相关文档数据库，然后结合用户查询，采用三次项集剪枝策略，挖掘初检相关文档数据库中含有用户查询词项的矩阵加权特征词关联规则，构建矩阵加权关联规则库。具体的剪枝策略是：第一次剪枝为比较候选 k_项集权值 $W(C_k)$和 $KIWT(k, k+1)$[17]，剪除其 $W(C_k) < KIWT(k, k+1)$的候选项集 C_k；第二次是挖掘到 2_项集时，剪除不含查询项的候选 2_项集 C_2，主要原因是本文检索模型只是挖掘与原查询相关的频繁项集和矩阵加权关联规则，而认为不含中文查询词项的候选 2_项集中的词项是与原查询不相关的，选择在候选 2_项集做删除处理是为了减少后续这类与原查询不相关的项集数量，提高挖掘效率；第 3 次是剪除其支持计数为 0 的候选项集 C_k。上述挖掘思想形式化为 MWARM_OQT(Matrix Weighted Association Rule Mining with Original Query Terms)算法。

输入：目标语言初检相关文档集(Doc^{TL})，最小支持度和置信度阈值[17](ms, mc)，印尼语用户查询(Q^{SL})。

输出：目标语言特征词矩阵加权关联规则集合($mwAR^{TL}$)。

Begin

let $mwFI^{TL}$←ϕ, $mwAR^{TL}$←ϕ;

//$mwFI^{TL}$ 为特征词矩阵加权频繁项集集合，$mwFI^{TL}$ 和 $mwAR^{TL}$ 清空。

(Doc^{TL}_DB)←Preprocessing(Doc^{TL});

//文档预处理模块对 Doc^{TL} 进行预处理，构建用户反馈初检相关文档数据库 Doc^{TL}_DB。本模型中，Doc^{TL} 是中文文档，其预处理包括分词、去停用词和提取中文特征词等。模型中所用的分词系统是中国科学院计算技术研究所研制编写的汉语词法分析系统 ICTCLAS①。

(C_1, $w(C_1)$, n_{c1}, $KIWT$ (1, 2))←ScanForC_1(Doc^{TL}_DB);

//扫描初检相关文档数据库 Doc^{TL}_DB，提取特征词 1_候选项集 C_1，计算 C_1 支持计数 n_{c1} 及其权值 $w(C_1)$ 和 $KIWT$ (1, 2)的值。$KIWT$ (1, 2)的计算公式见文献[17]

L_1←{C_1| $mwsupport(C_1)$≥ms};

//从 1_候选项集 C_1 挖掘 1_频繁项集，$mwsupport(C_1)$为 C_1 的矩阵加权支持度，$mwsupport(C_1)=w(C_1)/n_{c1}$ [17]。

for (k=2; $C_k \neq \phi$, k++){

//挖掘含有查询项的矩阵加权频繁 k_项集(k≥2)

$mwFI^{TL}$←$mwFI^{TL}$ ∪L_{k-1};

//频繁项集添加到 $mwFI^{TL}$ 集合

C_{k-1}←FirstPruning($w(C_{k-1})$, $KIWT(k-1, k)$);

//比较候选项集权值和 $KIWT$ 值，剪除其 $W(C_{k-1})$<$KIWT(k-1, k)$的候选项集 C_{k-1}, $KIWT(k-1, k)$的计算公式见文献[17]

C_k←CJoin (C_{k-1}); //候选项集 C_{k-1} 进行 Aproiri 连接[18]，得到 C_k

if (k=2)then C_k←SecondPruning(C_k , Q^{SL});

//挖掘到 2_项集时，剪除不含查询项的候选 2_项集

($w(C_k)$, n_{ck}, $KIWT$ (k, k+1))←ScanForC_k(Doc^{TL}_DB);

//扫描初检相关文档数据库 Doc^{TL}_DB，统计 C_k 的支持计数 n_{ck}，计算 C_k 权值 $w(C_k)$和 $KIWT$ (k, k+1)的值。$KIWT$ (k, k+1) 的计算公式见文献[17]

C_k←ThirdPruning(C_k);//剪除 n_{ck} 为 0 的候选项集 C_k;

L_k←{ C_k | $mwsupport(C_k)$≥ms};

//从 k_候选项集 C_k 挖掘 k_频繁项集，$mwsupport(C_k)$为 C_k 的矩阵加权支持度，$mwsupport(C_k)= w(C_k)/(n_{ck}\times k)$ [17]

}

for $mwFI^{TL}$ 中每一个频繁项集 I^{TL} do

//挖掘特征词矩阵加权关联规则

{

for I^{TL} 中每一对子项集 I_1 和 I_2 do

{

if ((I_1∪ I_2=I^{TL})and (I_1∩I_2= ∅)) then

{

计算 $mwconf(I_1→I_2)$和 $mwconf(I_2→I_1)$的值;

// $mwconf(I_1→I_2)$和 $mwconf(I_2→I_1)$为关联规则的置信度

if $mwconf(I_1→I_2)$≥mc

then $mwAR^{TL}$←$mwAR^{TL}$∪{ I_1→I_2};

if $mwconf(I_2→I_1)$≥mc

then $mwAR^{TL}$←$mwAR^{TL}$∪{ I_2→I_1};

}

}

}

output($mwAR^{TL}$); //输出含有查询项的矩阵加权强关联规则

End

①http://ictclas.nlpir.org/.

其中，关联规则的置信度计算公式[17]如下。

$$mwconf(I_1 \rightarrow I_2)= mwsupport(I_1, I_2)/mwsupport(I_1) \quad (1)$$

$$mwconf(I_2 \rightarrow I_1)= mwsupport(I_1, I_2)/mwsupport(I_2) \quad (2)$$

(2) 检索模型中印尼中跨语言查询扩展模型

本文检索模型中，其译后查询扩展词的来源是上述 MWARM_OQT 算法对目标语言初检用户相关文档集挖掘得到的矩阵加权关联规则，这些规则的前件是译后目标语言原始查询词项集合(Q^{TL})，而规则的后件是目标语言扩展词项目集合(ET^{TL})，通过矩阵关联规则的置信度 $mwconf$ 值确定了查询词项与扩展词项的关联程度。因此，其跨语言查询扩展模型(Cross Language Query Expansion Model, CLQEM)描述如公式(3)所示:

$$\text{CLQEM}=(Q^{TL},\ ET^{TL},\ W_q,\ W_{ET}) \qquad (3)$$

其中，

$$\begin{cases} Q^{TL}=\{q_1, q_2, \cdots, q_n\},\ q_n(n\geq 1)\text{为查询词项} \\ ET^{TL}=\{t_1, t_2, \cdots, t_m\},\ t_n(m\geq 1)\text{为扩展词项} \\ Q^{TL} \rightarrow ET^{TL}(mwsupport\geq ms,\ mwconf\geq mc) \\ W_q=(0.5+\dfrac{0.5\times tf_q}{\max(tf_q)})\times\log\dfrac{N}{df_q}\ ^{[19]} \end{cases}$$

$$W_{ET}=\max(mwconf)$$

在上述扩展模型中，W_q 表示译后原查询 Q^{TL} 的查询项 q 权值，tf_q 为查询项 q 在查询中的初始频率，$\max(tf_q)$表示所有查询项初始频率中的最高者，df_q 为包含查询项 q 的初检文档数，N 为初检相关文档总数。W_{ET} 表示来自矩阵关联规则 $Q^{TL} \rightarrow ET^{TL}$ 的目标语言查询扩展词权值，其值等于矩阵关联规则的置信度值。W_{ET} 表达式表明当扩展词重复出现在不同的矩阵关联规则时，就会存在不同的置信度，取其置信度最高者作为该扩展词权值。

(3) 基于矩阵加权关联模式挖掘的印尼中跨语言信息检索算法

在本文跨语言检索模型中，基于矩阵加权关联模式挖掘的印尼中跨语言信息检索基本思想是：采用跨

语言两次检索策略, 首先将印尼语查询通过机器翻译系统译为中文查询, 并提交搜索引擎在互联网中检索中文文档, 通过用户点击下载行为获取跨语言用户反馈初检相关文档集, 调用 MWARM_OQT 算法对用户反馈初检相关文档集进行挖掘, 得到与原查询相关的矩阵加权关联规则, 从关联规则中提取扩展词实现跨语言查询译后扩展, 将扩展词和原查询组合为新查询再次提交搜索引擎检索中文文档, 得到的最终检索结果通过机器翻译系统译为印尼语文档返回给用户。上述思想形式化为 ICCLIR_MWAR (Indonesian-Chinese Cross Language Information Retrieval Based on Matrix-Weighted Association Rules)算法。

输入: 印尼语用户查询(Q^{SL}), 最小支持度和置信度阈值(*ms*, *mc*)。
输出: 查询扩展后的跨语言检索结果(印尼语文档和中文文档)。
Begin
Q^{TL}←ExecMTranslate (Q^{SL});
//将印尼语用户查询 Q^{SL}(即源语言查询)提交给机器翻译系统(Microsoft translator API), 经过翻译后得到中文查询 Q^{TL}(即目标语言查询), 采用 ICTCLAS 系统完成译后中文查询 Q^{TL} 预处理。
FirstRDoc←FirstRetrieval (Q^{TL}, W_q);
//将翻译后的中文查询提交给搜索引擎, 如百度或谷歌等, 通过互联网检索中文文档, 得到跨语言初检结果中文文档集。
Doc^{TL}←UserClickDowdload (*FirstRDoc*);
// 根据用户浏览初检结果文档集 *FirstRDoc* 的点击、浏览、下载行为, 构建用户反馈初检相关文档集 Doc^{TL}。(如果存在用户对初检文档的点击下载行为, 则认为该篇文档与原查询是相关的, 应该从初检文档集中提取该篇文档)。
$mwAR^{TL}$ ←MWARM_OQT (Doc^{TL} , *ms*, *mc*, Q^{TL});
//调用用 MWARM_OQT 算法挖掘目标语言特征词矩阵加权关联规则 $mwAR^{TL}$, 并构建规则库。
(ET^{TL}, W_{ET})←Get Exp_Term ($mwAR^{TL}$);
//从 $mwAR^{TL}$ 集合中提取目标语言扩展词 ET^{TL}, 根据公式(2)计算扩展词权值 W_{ET}。
TL_Doc←SecondRetrieval (Q^{TL}, ET^{TL});
//将原查询和扩展词组合再次在互联网中检索目标语言文档, 得到最终目标语言文档 *TL_Doc*, 即中文文档。
SL _Doc←ExecMTranslate (*TL_Doc*);
//将目标语言文档 *TL_Doc*(中文文档)机器翻译为源语言文档 *SL _Doc*(印尼语文档)。
outputToUser (*TL_Doc*, *SL _Doc*);
//将查询扩展后检索结果中文文档和印尼文档返回给用户。
End

3 实验设计及其结果分析

根据上述理论分析和所给的模型结构图, 编写基于向量空间模型和矩阵加权关联模式挖掘的印尼中跨语言信息检索模型源程序进行实验。实验的硬件环境是: Intel(R) Core(TM) i7-3770 CPU @3.4GHz 3.4GHz 台式电脑, 内存 8.0GB, 硬盘 1TB; 软件环境为: Windows 7+VC#+SQL Server。

3.1 数据集及其预处理

采用日本情报信息研究所主办的多国语言处理国际评测会议上的跨语言信息检索标准数据测试集 NTCIR-5 CLIR①的 Economic Daily News 2000 年中文新闻文本作为本实验语料, 共计 79 380 篇中文文本信息。NTCIR-5 CLIR 有查询集、文档测试集以及结果集。其中, 查询集有 50 个查询主题, 分有 TITLE、DESC、NARR 和 CONC 等 4 种类型, 本文实验选择 TITLE 和 DESC 类型, TITLE 类型查询主题以名词和名词性短语简要描述, 属于短查询; DESC 类型以句子形式简要描述查询主题, 属于长查询。其结果集有 Rigid 和 Relax 等两种评价标准, Rigid 标准是指其答案都是与原查询相关或高度相关的; Relax 标准是指高度相关、相关或部分相关的。

为了进行本文印尼中跨语言信息检索模型的实验, 邀请翻译机构的专业翻译人士先将 NTCIR-5 CLIR 中文版 50 个查询主题人工翻译为印尼语, 再进行查询。

3.2 基准实验及其实验评价指标

为了验证本文提出的印尼中跨语言信息检索模型的有效性, 选择中文单语言检索(Monolingual Retrieval Baseline, MRB)和没有查询扩展的印尼中跨语言检索(Cross-language Retrieval Baseline, CLRB), 以及传统的基于伪相关反馈的印尼中跨语言信息检索算法[2] (Cross-Language Retrieval Using Pseudo Relevance Feedback, CLR_PRF)作为实验基准, 与本文检索模型的检索性能进行比较和分析。

上述三种基准的检索结果是: MRB 基准是用中文查询直接检索中文文档得到的检索结果; CLRB 是印尼查询经机器翻译系统翻译为中文查询检索中文文档得到的检索结果, 即传统的跨语言信息检索结果; CLR_PRF 基准是在如下参数设置下实现跨语言查询扩展后再次检索得到的结果, 其参数设置(与文献[2]

①http://research.nii.ac.jp/ntcir/permission/ntcir-5/perm-en-CLIR.html.

一致)是：提取跨语言前列初检文档 20 篇构建初检相关文档集，提取前列权值(降序排列)的 20 个特征词为扩展词。

采用 R-查准率(R_prec)、P@10 和 P@20 作为实验评价指标。R-查准率(R-prec)是当 R 个文档被检索后所计算的查准率，其中 R 是指对应于某个查询在文档集合中相关文档数，不强调文档结果集中文档的排序情况，由于 NTCIR-5 CLIR 测试集中不同查询主题的相关文档数差别比较大，故该指标值显得更有意义和评价价值。

3.3　实验结果及其分析

运行本文检索模型源程序，将该模型与基准算法 MRB、CLRB 和 CLR_PRF 在 NTCIR-5 CLIR 测试集上进行文本检索，对其检索性能进行比较和分析。同时，分析矩阵加权支持度和置信度参数对本文模型检索性能的影响。

(1) 基准实验结果及分析

为了与本文检索模型的检索性能比较，先运行 MRB、CLRB、CLR_PRF 等三个基准源程序，提交 NTCIR-5 CLIR 的 50 个查询主题的 TITLE 和 DESC 部分的中文查询进行中文单语言检索基准实验，以及印尼语查询进行印尼中跨语言检索和传统的基于伪相关反馈的印尼中跨语言检索基准实验，得到基准实验结果，如表 1 所示。

表 1　三种基准算法跨语言检索实验结果

查询类型	评测类型	评价指标	MRB	CLRB	CLRB 占 MRB (%)	CLR_PRF	CLR_PRF 占 MRB (%)	CLR_PRF 比 CLRB 提高(%)
TITLE	Relax	R_prec	0.258	0.1313	50.89	0.1278	49.53	–2.67
		p@10	0.2292	0.0792	34.55	0.1083	47.25	36.74
		p@20	0.1542	0.0625	40.53	0.0792	51.36	26.72
	Rigid	R_prec	0.1919	0.1442	75.14	0.1113	58.00	–22.82
		p@10	0.1417	0.0458	32.32	0.0625	44.11	36.46
		p@20	0.0979	0.0333	34.01	0.0479	48.93	43.84
DESC	Relax	R_prec	0.227	0.1205	53.08	0.0354	15.59	–70.62
		p@10	0.2375	0.1333	56.13	0.0958	40.34	–28.13
		p@20	0.1667	0.1	59.99	0.0979	58.73	–2.10
	Rigid	R_prec	0.1867	0.1226	65.67	0.0587	31.44	–52.12
		p@10	0.15	0.0542	36.13	0.0458	30.53	–15.50
		p@20	0.1063	0.0458	43.09	0.0521	49.01	13.76

从表 1 可以看出，传统的跨语言检索 CLRB 基准只达到了单语言检索基准 MRB 的 32.32%至 75.14%，而传统的基于伪相关反馈的印尼中跨语言信息检索 CLR_PRF 检索效果更差，才达到了单语言基准 MRB 的 15.59%至 58.73%。与 CLRB 基准比较，CLR_PRF 检索结果的各个评价指标值中，大多数比 CLRB 检索结果的指标值减少了，减少幅度最大为 70.62%(即 DESC 类查询、Relax 评测类型的 R_prec 值)；只有少数指标值有所增加，提高幅度最大是 p@20 指标(TITLE 类查询、Rigid 评测类型)，达到 43.84%。

表 1 实验结果表明，印尼中跨语言基准(即传统的跨语言检索)的检索性能明显地低于单语言的基准检索性能，有些指标值最低只达到 15.59%。说明在传统的跨语言信息检索中，印尼查询经过机器翻译为中文查询后，受查询翻译质量的影响，查询主题漂移比较严重，即其检索出的相关文档比较少，而与查询非相关的文档比较多。而在查询主题漂移如此严重的情况下进行伪相关反馈查询扩展的跨语言检索，导致其检索性能更差，因此，CLR_PRF 的检索性能不如 CLRB 好。

(2) 本文跨语言检索模型与基准算法的检索性能比较

运行本文模型源程序，提交 NTCIR-5 CLIR 的 50 个查询主题的 TITLE 部分和 DESC 部分的印尼语查询进行印尼中跨语言检索实验，在支持度变化和置信度变化两种情况下与上述 3 个基准(MRB、CLRB 和 CLR_PRF)进行检索性能比较和分析，其检索结果的 R_prec、p@10 和 p@20 值分别如表 2 和表 3 所示。本文的模型实验参数设置如下：提取跨语言初检文档前

列 100 篇文档提交给用户，用户进行点击、浏览、下载等行为后确定初检相关文档。为了实验方便，将初检前列中含有已知结果集的 100 篇相关文档视为用户在点击、浏览后，下载的相关反馈文档信息。另外，所挖掘的项集长度为 3，支持度变化时的实验参数为置信度 *mc*=0.01，支持度 *ms* 分别为 0.5、0.55、0.6、0.65、0.7 和 0.75 时得到检索结果的 R_prec、p@10 和 p@20 值，取平均值作为其在表 2 的值，置信度变化时的实验参数：支持度 *ms*=0.5，置信度 *mc* 分别为 0.008、0.01、0.05、0.08 和 0.1 时得到结果如表 3 所示。

表 2　支持度变化时本文检索模型与基准算法的检索性能比较

查询类型	评测类型	评价指标	本文检索模型	本文模型占 MRB (%)	本文模型比 CLRB 提高(%)	本文模型比 CLR_PRF 提高(%)
TITLE	Relax	R_prec	0.2355	91.28	79.36	84.27
		p@10	0.1410	61.52	78.03	30.19
		p@20	0.1056	68.46	68.91	33.33
	Rigid	R_prec	0.2176	113.39	50.90	95.51
		p@10	0.0903	63.70	97.09	44.48
		p@20	0.0653	66.67	96.00	36.33
DESC	Relax	R_prec	0.2383	104.99	97.79	573.16
		p@10	0.1882	79.24	41.19	96.45
		p@20	0.1424	85.41	42.38	45.45
	Rigid	R_prec	0.2321	124.32	89.31	295.40
		p@10	0.0896	59.72	65.28	95.63
		p@20	0.0764	71.87	66.81	46.64

表 3　置信度变化时本文检索模型与基准算法的检索性能比较

查询类型	评测类型	评价指标	本文检索模型	本文模型占 MRB (%)	本文模型比 CLRB 提高(%)	本文模型比 CLR_PRF 提高(%)
TITLE	Relax	R_prec	0.2351	91.14	79.09	83.99
		p@10	0.1392	60.72	75.73	28.51
		p@20	0.1021	66.21	63.36	28.91
	Rigid	R_prec	0.2433	126.78	68.72	118.60
		p@10	0.0867	61.16	89.21	38.66
		p@20	0.0633	64.70	90.21	32.23
DESC	Relax	R_prec	0.2295	101.09	90.44	548.25
		p@10	0.1842	77.55	38.17	92.25
		p@20	0.1371	82.23	37.08	40.02
	Rigid	R_prec	0.2133	114.24	73.96	263.34
		p@10	0.0942	62.77	73.73	105.59
		p@20	0.0767	72.14	67.42	47.18

从表 2 实验结果可知，当支持度变化时，本文检索模型检索结果的各个评价指标值是单语言检索基准 MRB 的 59.72%(最低)至 124.32%(最高)范围，比跨语言基准算法 CLRB 检索结果的各个指标值提高 41.19%(最低)至 97.79%(最高)范围；比基于伪相关反馈的印尼中跨语言检索基准 CLR_PRF 的提高 30.19%(最低)至 573.16%(最高)范围，效果比较显著。另外，表 2 还表明，长查询类型 DESC 的检索效果比短查询类型 TITLE 的好，对于长查询类型 DESC，本文检索模型检索结果的 Rigid 类型的 R_prec 值比单语言检索的提高了 24.32%(即(0.2321–0.1867)/0.1867)。

表 3 实验结果表明，当置信度阈值变化时，本文检索模型检索结果的各个评价指标值占单语言检索基准 MRB 的 60.72%至 126.78%范围，最好的情况是其

长查询类型DESC的R_prec值比单语言检索的提高了14.25%(即 Rigid 类型的 R_prec 值: (0.2133–0.1867)/0.1867)。与跨语言基准算法 CLRB 比较, 本文检索模型检索结果的各个评价指标值提高 37.08%至 90.44%, 同时, 比 CLR_PRF 基准的提高了 28.51%至 548.25%, 效果比较显著。另外, 表 3 还表明, 长查询类型 DESC 的检索效果比短查询类型 TITLE 的好。

(3) 支持度和置信度对本文模型的检索性能影响

在不同的矩阵加权支持度阈值 *ms* 和置信度阈值 *mc* 下, 本文印尼中跨语言检索模型检索性能如表4(其中矩阵加权置信度 *mc*=0.01)和表 5(其中矩阵加权支持度 *ms*=0.5)所示。

表 4 支持度变化时本文跨语言检索模型的检索性能(*mc*=0.01)

查询类型	评测类型	评价指标	矩阵加权支持度 *ms*					
			0.5	0.55	0.6	0.65	0.7	0.75
TITLE	Relax	R_prec	0.2359	0.2361	0.234	0.2328	0.2318	0.2424
		p@10	0.1417	0.1625	0.1417	0.1417	0.1417	0.1167
		p@20	0.1042	0.1104	0.1021	0.1021	0.1000	0.1146
	Rigid	R_prec	0.2443	0.2443	0.2032	0.202	0.2008	0.211
		p@10	0.0875	0.1083	0.0875	0.0875	0.0875	0.0833
		p@20	0.0646	0.0708	0.0625	0.0625	0.0604	0.0708
DESC	Relax	R_prec	0.2399	0.2376	0.2367	0.2371	0.2332	0.2455
		p@10	0.1875	0.1917	0.1792	0.1875	0.1875	0.1958
		p@20	0.1396	0.1438	0.1458	0.1438	0.1396	0.1417
	Rigid	R_prec	0.2443	0.2421	0.2413	0.242	0.2056	0.2173
		p@10	0.0958	0.0917	0.0875	0.0875	0.0833	0.0917
		p@20	0.0771	0.0771	0.0792	0.0771	0.0729	0.075

表 5 置信度变化时本文跨语言检索模型的检索性能(*ms* =0.5)

查询类型	评测类型	评价指标	矩阵加权置信度 *mc*				
			0.008	0.01	0.05	0.08	0.1
TITLE	Relax	R_prec	0.2362	0.2359	0.2349	0.2345	0.2342
		p@10	0.1417	0.1417	0.1417	0.1375	0.1333
		p@20	0.1042	0.1042	0.1021	0.1	0.1
	Rigid	R_prec	0.2445	0.2443	0.2434	0.2425	0.2418
		p@10	0.0875	0.0875	0.0875	0.0875	0.0833
		p@20	0.0646	0.0646	0.0625	0.0625	0.0625
DESC	Relax	R_prec	0.2399	0.2394	0.2401	0.2156	0.2124
		p@10	0.1875	0.1875	0.1875	0.1792	0.1792
		p@20	0.1396	0.1375	0.1396	0.1354	0.1333
	Rigid	R_prec	0.2443	0.1402	0.2444	0.2204	0.2171
		p@10	0.0958	0.0958	0.0958	0.0917	0.0917
		p@20	0.0771	0.0771	0.0771	0.0771	0.075

从表 4 和表 5 可以看出, 对于 TITLE 和 DESC 类型查询, 随着矩阵加权支持度或置信度阈值的不断提高, 本文检索模型检索结果的 R_prec、p@10 和 p@20 值变化比较缓慢, 有些呈现下降的趋势。主要原因分析如下: 在查询主题严重漂移的情况下, 随着矩阵加权支持度或置信度阈值的不断提高, 从矩阵加权词间关联规则中获得的扩展词逐渐减少, 导致跨语言检索性能下降; 反之, 当支持度或者置信度阈值下降时,

检索系统获得的扩展词会多些，跨语言检索性能得到改善和提升。但是，当扩展词增多时，虚假的扩展词即噪音出现的机会也增多，此时也会导致检索性能降低。因此，如何确定一个合适的支持度或置信度阈值，是值得研究的问题。

(4) 实验结果分析

理论分析和实验结果表明，与单语言检索基准MRB、传统的跨语言检索基准CLRB和传统的基于伪相关反馈的跨语言查询算法CLR_PRF比较，本文提出的印尼中跨语言检索模型能有效地减少查询主题漂移问题，其检索性能获得了很大的改善和提高。表2和表3实验结果表明，其检索结果的R_prec、p@10和p@20值均达到单语言检索基准MB的60%以上，最好的情况是其R_prec值比单语言检索提高了24.32%。特别地，其检索结果比跨语言检索基准CLRB和CLR_PRF的好，提高最大幅度达到548.25%。这些实验结果表明，本文提出的印尼中跨语言信息检索模型是有效的，能改善和提高跨语言信息检索性能。其主要原因分析如下：在跨语言信息检索中，查询翻译结果对跨语言检索结果影响较大，常常导致跨语言初检结果质量不如单语言的初检结果，即出现严重的查询主题漂移问题，而将用户浏览、点击、下载行为，矩阵加权关联模式挖掘与查询扩展等技术融合应用到印尼中跨语言信息检索模型，可以获得与原查询最相关的反馈信息，通过矩阵加权关联规则挖掘得到与原查询相关的扩展词实现跨语言查询扩展，可极大减少跨语言检索中存在的严重主题漂移问题，提高印尼中跨语言检索性能。

同时，矩阵加权支持度和置信度对本文的印尼中跨语言信息检索模型的检索性能是有影响的，矩阵加权支持度或置信度过高，会遗漏一些与原查询相关的扩展词，导致跨语言查询扩展性能降低；反之，如果其过低，与原查询不相关的扩展词会出现或增多，严重的情况会导致新的查询主题漂移。因此，如何取得一个合适的支持度和置信度阈值是值得研究的课题。

4 结 语

随着中国和东盟国家各个领域的交流日益加深，针对东盟国家语言的跨语言信息检索与跨语言信息服务研究显得迫切和重要。本文以印尼语和汉语为研究对象，将用户点击行为与矩阵加权关联模式挖掘融合引入印尼中跨语言信息检索模型，阐述了该模型实现的关键技术，实验结果表明，本文所提的模型是有效的，能减少查询主题漂移，解决了跨语言信息检索长期存在的严重主题漂移问题，提高和改善印尼中跨语言信息检索性能，对长查询的检索效果更好。

由于搜索引擎的研究范围广以及要考虑的因素比较多，本文的实验工作是在基于向量空间模型的跨语言检索系统中进行的，是模拟实验。下一步研究重点是：将该检索模型实用化，开发搜索引擎环境下实用的印尼中跨语言信息检索系统，同时，深入研究矩阵加权关联模式挖掘参数对印尼中跨语言检索性能的影响，找出其变化的规律，以便推广到实际系统中。

(致谢：感谢匿名外审专家以及编辑部的修改意见。)

参考文献：

[1] Gao J F, Nie J Y, Zhang J, et al. TREC-9 CLIR Experiments at MSRCN [C]//Proceedings of the 9th Text Retrieval Evaluation Conference. 2001.

[2] 吴丹，何大庆，王惠临. 基于伪相关反馈的跨语言查询扩展[J]. 情报学报，2010，29(2)：232-239. (Wu Dan, He Daqing, Wang Huilin. Cross-Language Query Expansion Using Pseudo Relevance Feedback [J]. Journal of the China Society for Scientific and Technical Information, 2010, 29(2): 232-239.)

[3] 吴丹，何大庆，王惠临. 一种基于相关反馈的跨语言信息检索查询翻译优化技术研究[J]. 情报学报，2012，31(4)：398-406. (Wu Dan, He Daqing, Wang Huilin. A Relevance Feedback Based Query Translation Enhancement Technique in Cross Language Information Retrieval [J]. Journal of the China Society for Scientific and Technical Information, 2012, 31(4): 398-406.)

[4] Chinnakotla M K, Raman K, Bhattacharyya P. Multilingual Pseudo-relevance Feedback: Performance Study of Assisting Languages [C]//Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2010: 1346-1356.

[5] Parton K, Gao J. Combining Signals for Cross-Lingual Relevance Feedback [C]//Proceedings of the 8th Asia Information Retrieval Societies Conference (AIRS 2012), Tianjin, China. Springer Berlin Heidelberg. 2012.

[6] Lee C J, Croft W B. Cross-Language Pseudo-Relevance Feedback Techniques for Informal Text [C]//Proceedings of

the 36th European Conference on IR Research (ECIR 2014), Amsterdam, The Netherlands. Springer International Publishing, 2014.

[7] 闭剑婷, 苏一丹. 基于潜在语义分析的跨语言查询扩展方法[J]. 计算机工程, 2009, 35(10): 49-50. (Bi Jianting, Su Yidan. Expansion Method for Language-crossed Query Based on Latent Semantic Analysis [J]. Computer Engineering, 2009, 35(10): 49-50.)

[8] 魏露, 李书琴, 李伟男, 等. 跨语言查询扩展优化[J]. 计算机工程与设计, 2014, 35(8): 2785-2788, 2803. (Wei Lu, Li Shuqin, Li Weinan, et al. Optimization of Cross-language Query Expansion [J]. Computer Engineering and Design, 2014, 35(8): 2785-2803.)

[9] 宁健, 林鸿飞. 基于改进潜在语义分析的跨语言检索[J]. 中文信息学报, 2010, 24(3): 105-111. (Ning Jian, Lin Hongfei. Cross-Language Information Retrieval Based on Improved Latent Semantic Indexing [J]. Journal of Chinese Information Processing, 2010, 24(3): 105-111.)

[10] 罗远胜, 王明文, 勒中坚, 等. 跨语言信息检索中的双语主题相关模型[J]. 小型微型计算机系统, 2013, 34(12): 2758-2763. (Luo Yuansheng, Wang Mingwen, Le Zhongjian, et al. Bilingual Topic Correlation Model in Cross-lingual Information Retrieval [J]. Journal of Chinese Computer Systems, 2013, 34(12): 2758-2763.)

[11] Rahimi R, Shakery A, King I. Multilingual Information Retrieval in the Language Modeling Framework [J]. Information Retrieval Journal, 2015, 18(3): 246-281.

[12] Ganguly D, Leveling J, Jones G J F. Cross-lingual Topical Relevance Models [C]//Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012). 2012.

[13] Wang X W, Zhang Q, Wang X J, et al. LDA Based PSEUDO Relevance Feedback for Cross Language Information Retrieval [C]//Proceedings of the 2nd International Conference on Cloud Computing and Intelligence Systems. IEEE, 2012.

[14] Wang X W, Wang X J, Zhang Q, et al. A Web-Based CLIR System with Cross-Lingual Topical Pseudo Relevance Feedback [C] // Proceedings of the 4th International Conference on Conference and Labs of the Evaluation Forum (CLEF) Initiative, Valencia, Spain. 2013.

[15] 王序文, 王小捷, 孙月萍. 双语主题跨语言伪相关反馈[J]. 北京邮电大学学报, 2013, 36(4): 81-84. (Wang Xuwen, Wang Xiaojie, Sun Yueping. Cross-lingual Pseudo Relevance Feedback Based on Bilingual Topics [J]. Journal of Beijing University of Posts and Telecommunications, 2013, 36(4): 81-84.)

[16] Wang X W, Zhang Q, Wang X J, et al. Cross-lingual Pseudo Relevance Feedback Based on Weak Relevant Topic Alignment [C]//Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation Shanghai, China. 2015: 529-534.

[17] 黄名选, 严小卫, 张师超. 基于矩阵加权关联规则挖掘的伪相关反馈查询扩展[J]. 软件学报, 2009, 20(7): 1854-1865. (Huang Mingxuan, Yan Xiaowei, Zhang Shichao. Query Expansion of Pseudo Relevance Feedback Based on Matrix-Weighted Association Rules Mining [J]. Journal of Software, 2009, 20(7): 1854-1865.)

[18] Agrawal R, Imielinski T, Swami A. Mining Association Rules Between Sets of Items in Large Database[C]//Proceedings of 1993 ACM SIGMOD International Conference on Management of Data. 1993.

[19] Salton G, Buckley C. Term-weighting Approaches in Automatic Text Retrieval [J]. Information Processing & Management, 1988, 24(5): 513-523.

利益冲突声明:

作者声明不存在利益冲突关系。

支撑数据:

支撑数据见期刊网络版 http://www.infotech.ac.cn。

[1] 黄名选. result.xls. 研究结果数据.

收稿日期: 2016-09-18
收修改稿日期: 2016-11-09

Cross Language Information Retrieval Model Based on Matrix-weighted Association Patterns Mining

Huang Mingxuan
(Guangxi Key Laboratory Cultivation Base of Cross-border E-commerce Intelligent Information Processing, Guangxi University of Finance and Economics, Nanning 530003, China)
(Department of Computer Science, Guangxi University of Finance and Economics, Nanning 530003, China)

Abstract: [**Objective**] The purpose of this paper is to solve the query drift issue facing cross language information retrieval. It proposes a new model to retrieve Chinese documents with Indonesian queries. [**Methods**] The new model integrated the algorithms of matrix-weighted association patterns mining, query expansion, as well as user click-download behaviors. [**Results**] The R_prec, p@10 and p@20 values of the proposed model were higher than the 60% benchmark of the monolingual retrieval on the CLIR NTCIR-5 data set. These results were 37% higher than cross language retrieval baseline and 28% higher than the existing algorithms based on pseudo relevance feedback. [**Limitations**] The proposed model was only examined in the cross language retrieval system built with the vector space model, which needs to be done with the real world search engines. [**Conclusions**] The proposed model could effectively reduce query drift in cross language retrieval, and retrieve more relevant Chinese documents with Indonesian long queries.

Keywords: Click Behavior   Association Patterns Mining   Indonesian-Chinese Cross Language Retrieval Model   Cross Language Information Retrieval   Matrix-weighted Association Rule

HighWire Press 收购 Semantico

学术出版公司 HighWire Press 于近日宣布成功收购 Semantico, Semantico 是一家为学术出版市场提供技术和服务的私企。这项收购将使得 HighWire 提高其技术创新能力, 团队变得更加强大, 产品组合变得更加丰富。

"创新和以客户为中心是 HighWire 的核心, Semantico 解决方案集和整个团队的加入, 提高了我们的产品服务能力, 有助于我们服务于整个行业。" HighWire CEO Dan Filby 说: "这次收购也符合我们公司的长期增长和价值创造战略。"

Semantico 董事长兼创始人 Richard Padley 补充: "我们的团队能够加入到 HighWire, 我感到非常兴奋。整合后更大规模的组织将有更强的服务能力, 将为当前和未来的出版商带来巨大的价值。"

HighWire 的创新解决方案包括:

(1) JCore: 行业领先、同类产品中最佳的开放式期刊平台;

(2) Folio: 针对学术研究的动态电子书平台;

(3) Scolaris: 针对多样化、专业化内容进行了优化的综合发布解决方案;

(4) SAMS Sigma: 基于云的、业界领先的、与访问管理集成的身份管理解决方案;

(5) BenchPress: 在线投稿和同行评议跟踪系统;

(6) Impact and Usage Vizors: 可视化分析工具, 提供无与伦比的洞察力, 为基于证据的出版决策提供支持;

斯坦福大学图书馆员、HighWire 董事会成员兼学术顾问 Mike Keller 表示: "HighWire 继续为客户提供更高的价值, 并且这次收购将有望进一步促进他们作为行业顶尖出版技术提供商的努力。"

(编译自: http://home.highwire.org/news/highwire-press-acquires-semantico)

(本刊讯)